Where's the Bias? Developing Effective Model Governance

Galen Harrison
Department of Computer Science
University of Chicago
Chicago, IL 60615
harrisong@uchicago.edu

Natasha Duarte
Center for Democracy and Technology
Washington, DC 20005
nduarte@cdt.org

Joseph Lorenzo Hall
Center for Democracy and Technology
Washington, DC 20005
joe@cdt.org

Abstract

Understanding and mitigating the discriminatory behavior of a machine learning algorithm and a system that uses a machine learning algorithm are related, but distinct tasks. We argue that regulatory procedures that are supposed to address discrimination in systems that use machine learning, but which only consider the algorithm or model will miss significant sources of discrimination. We support this argument by considering applications of machine learning to financial services, and showing how bias could arise both through areas of specific machine learning concern and broader issues of design. Because bias may extend from both these concerns, we argue that an effective model governance regime cannot limit itself to scrutinizing the model.

1 Introduction

As industry has started to think about using machine learning to fulfill business functions, there has been an increasing interest in how to perform model governance (Burt et al., 2018). Model governance is the process by which one oversees the development and use of systems that use machine learning models, specifically with the purpose of ensuring that the system does not have any undesirable effects. While initially meant to address the appropriateness of risk calculations for financial institutions (Ordóñez-Sanz, 2014), it has been expanded to a more broad consideration of quantitative methods. At the forefront of the effects that governance is concerned with is bias in some form or another. The issue of governance and the promises of machine learning come together in the area of finance. Because machine learning has such a potential to reduce friction and increase access, simply not using it is not a tenable situation. However, the same factors that create this impetus, also increase the consequences of bias. Access to financial products can radically alter the trajectory of someone's life, so a system that discriminates unfairly can have a significant negative impact. Therefore, model governance should be of the utmost concern.

Any model governance procedure that has been formulated to prevent prejudice needs to have an understanding of bias. Bias in machine learning is, and has been treated as, a machine learning problem. That is to say, attempts have been made to translate what it means to be biased into a language that is tractable to optimization, classification and probability. But bias is not just a machine

NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, Montréal, Canada. This is a non-archival publication - the authors may submit revisions and extensions of this paper to other publication venues.

learning problem, it may manifest outside of any model or learning setting. Bias that arises in a system that uses machine learning may occur due to decisions to target a specific audience, or to use particular forms of information, neither of which have anything to do with the essential machine learning problem. However, because ML may enable these biases to occur at a large scale and with much less human assistance, they must also be considered within a model governance framework.

Why should broader issues of bias in computer systems be the concern of someone interested in model governance? ML systems take less human intervention, which means that biased outcomes that might be noticed by a person will be less likely to be flagged. So ML, in a different sense, enables discrimination. But this connection aside, the ultimate purpose of this line of work is to prevent the systems that we study and help build from harming people, and artificially cutting off consideration of how ML may facilitate bias will hinder that goal. It's unlikely that those harmed by ML systems will particularly care to distinguish between discrimination caused by the model and that caused by the broader system components, so a governance process that limits itself to looking at the model will fail to actually prevent harm.

The rest of this paper is organized along the following lines. In section 2 we compare approaches to understanding bias in machine learning. In section 3, we identify five different financial tasks for which machine learning may or is being used. For each of these tasks, we describe a scenario in which there is bias that arises from machine learning concerns and a scenario in which bias arises from factors external to the machine learning aspect. The scenarios detailed are not by any means comprehensive, and aren't meant to be. Rather by providing materialized examples of biased outcomes, we hope to provide a model for the sorts of lines of inquiry for the system designers. We then conclude with several questions that ML governance entities should ask to help identify and address bias.

2 Frameworks for Understanding Bias

While some attention has been given to the legal dimensions of discriminatory behavior in machine learning (Barocas and Selbst, 2016; Baker and Dellaert, 2017), the goal of this paper is to examine potential ways in which machine learning applications may exhibit unfair bias that is not necessarily illegal bias. Some of the patterns of bias we consider may currently produce some form of legal liability, but as machine learning becomes more widespread, it's likely that governmental bodies will become increasingly interested in developing governance procedures. As legislative bodies take this cause up, it will be useful for the academic community have a theory of what precisely causes bias in machine learning.

There are a number of approaches towards understanding what precisely constitutes bias (or unfairness) in machine learning. There is a school of thought that has approached the task as a machine learning problem - posing the problem in probabilistic terms, and proposing and analyzing various methods for solving the problem in practice. Dwork et al. (2012), Joseph et al. (2016) Kilbertus et al. (2017), Feldman et al. (2015) Hardt et al. (2016) Hajian et al. (2012) are examples from this school of thought.

To understand how this school of thought approaches the problem, it is instructive to consider the appendix to "Fairness through Awareness" by Dwork et al. (2012), which contains a "catalog of evils". These are undesirable behaviors that a model may have, and which a mathematical definition of fairness should prohibit. They include classifying individuals in a minority group negatively at a higher rate, using a proxy variable to negatively classify likely minority group individuals (redlining), or randomly rejecting majority group individuals to mathematically rebut charges of unfavorable treatment. What's important to this analysis is that these definitions of fairness, while inspired by real world discrimination, are centered around the machine learning model, and concern specifically what the machine learning model is doing.

In contrast "Bias in Computer Systems" by Friedman and Nissenbaum Friedman and Nissenbaum (1996) provides an account of the ways in which computer systems (not ML specifically) can exhibit bias. The authors define bias as a systematic and unfair preference towards or against a particular group. Because fairness is situational and ultimately subjective, it's important to understand that the framework they develop is agnostic as to what actually constitutes bias, the framework they develop simply identifies situations where a preference for a certain group may be expressed. Ultimately for Friedman and Nissenbaum, bias is determined by the context in which the computer system is used.

They provide the example of a flight reservation system that inadvertently prioritized certain kinds of flights because of how the flights were ranked and the size of the screen they were displayed on.

It should be emphasized that nothing about the Friedman and Nissenbaum approach to bias contradicts the approach of viewing bias as a machine learning problem. Rather, the Friedman and Nissenbaum approach looks at specific biased outcomes that can arise. The scenarios specified in the "catalog of evils" could lead to any of the situations Friedman and Nissenbaum describe.

It is also important to note that looking into the model, either through trying to build explainability in, or induce particular fairness criteria during model construction, is a necessary but not sufficient component for understanding ML systems' impacts on the world (Kitchin, 2017; Ananny and Crawford, 2018). Prior work in this area has made specific assumptions about what discrimination may look like in machine learning systems, but as is suggested by the Friedman and Nissenbaum framework, discriminatory behavior can stem from many different components in a ML system. Particular model behaviors which may not look like bias when looking only at the model, may constitute bias when considered in a broader context.

While looking towards the future, there is also a need for governance in the present. The Friedman and Nissenbaum framework suggests that a governance process that is primarily concerned with testing the model, and examining the response to certain inputs will miss important aspects of biased behaviors. An effective model governance regime will need to engage in a holistic qualitative analysis of the systems' aims, values, imagined and actual audience, design process, and implementation. In the rest of this paper, we use examples from financial technology to illustrate how a machine learning understanding of bias can illuminate and address some, but not all, possible aspects of bias.

3 Financial Technology and Bias

FinTech is an umbrella term referring to a wide array of computational technologies meant to assist financial activities on the part of banks, consumers, or businesses. Startups and large banks are the main parties advancing the use of FinTech applications. The particular applications can range from advice (like SmartWallet) to ML for market prediction. In this paper we divide FinTech applications for which ML may be a key component into five categories: chatbots, advice and prediction, credit, fraud and identity, and product marketing. Financial technology is an instructive setting for considering bias because most adults in developed nations will have at least some form of contact with at least some forms of financial institutions. Furthermore, financial decisions can have a profound effect on the lives of consumers. In addition to being a profound and ubiquitous setting, it is also noteworthy that the tasks we identified span a wide range in terms of the machine learning task, and the machine learning task's role in the wider system. Thus, we expect that lessons learned in this domain will also apply to other domains in which machine learning may be applied.

In order to provide a more specific set of learning tasks, we surveyed the current state of startups along with publicly available information about ML use in large banks. While not necessarily empirical, we were systematic. We looked at the most recent startups with more than $1 million in funding that appear on Angel List when searching for either "finance" or "banking" (angellist.com, [n. d.]). We also searched for the top funded startups. For each category, we examined the top 30 startups that were not related to crypto-currencies. For each listed startup, we attempted to identify the functions that the startup claimed to perform (this was not always possible). We also looked at the copy associated with the largest american financial institutions: J.P. Morgan Chase, Bank of America, Wells Fargo, and Citigroup. From these materials, we derived a set of FinTech related tasks to which ML could provide core functionality.

3.1 Marketing

Marketing financial products, though not directly a financial service, is a FinTech function. Possibly out of a desire for vertical integration, Citigroup has invested in a few different startups which either facilitate customer data use for marketing purposes or which actually market financial products (Citi Ventures, [n. d.]). JP Morgan has also mentioned this as a possible avenue (though it's mentioned more in the context of trend forecasting) (J.P. Morgan, [n. d.]). Like credit assessment, marketing or recommendation systems are a well known setting for considering bias within the academic literature. A common way for targeted advertising to work is for platforms to sell particular audiences to people

trying to market their products. That is they will select combinations of things like “Likes the Rolling Stones". While financial institutions or financial apps can look at transaction and customer data and what financial products these customers did or did not purchase and make predictions about that, if they want to attract new customers rather than upselling existing ones, they will need to engage with external targeted advertising.

Speicher et. al. describe a machine learning framework for understanding bias in this setting (Speicher et al., 2018). One issue with using audience attributes is redlining. Specific interests or combinations of interests may correlate well with a sensitive attribute. The authors go on to identify methods by which an advertiser may use these sorts of combinations in different ad settings to create discriminatory ads. It’s worth noting that nothing about this setting necessarily uses “machine learning" in the sense that there doesn’t need to be any sort of optimization. If machine learning were used to determine which audiences to target with particular, it would exacerbate questions of opacity and agency, and would not mitigate any of the concerns raised by Speicher et. al. Even though “machine learning" as it is generally construed is not necessary for the approach taken in this paper, it is still an example of the machine learning approach simply because it focuses explicitly on the statistical relationship between a targeted group and the attributes of that group.

Using custom audiences is not necessarily the only way in which one might use machine learning to market products. While in the scenario considered by Speicher et. al., the audiences and who belongs to which audience are determined by the advertising platform (i.e. Facebook). A FinTech company with sufficient analytics capabilities may seek to develop their own categories. These categories could then be used for biased targeting. One could certainly use Speicher et. al.’s framework to examine how the actual targeting decisions are biased, but only if the governance process covers the targeting decisions. If the governance process is focused solely on the actual model, and isn’t scoped to include how the model is actually used then it will miss the bias.

3.2 Fraud

Fraud prevention and identity verification have also sparked some interest on the part of some larger banks along with a few startups. Fraud detection tasks range from detecting fraudulent transactions, to what are called KYC (know your customer) and AML (anti-money laundering) tasks. KYC and AML are duties that some financial institutions are required to fulfill to prevent their accounts from being used for illegal purposes. ML in fraud detection comes down to pattern recognition - identifying when transactions or account characteristics are atypical or have indicators of misuse. ML can also be used for identity verification (or at least to aid the process) through similar sorts of tasks. There are a variety of tasks that this domain can imply. Bolton and Hand suggest that there is a difference between identifying credit card or payment fraud (Bolton and Hand, 2002), and money laundering. They note that, while payment fraud can best be captured by looking either for anomalies or known fraudulent patterns, money laundering is generally addressed by link analysis. That is, the task is to group money transfers together in ways that are indicative of laundering.

There’s some interest in this domain from larger banks, as well as some startups. Citigroup has invested in startups which may use ML for fraud detection (Citi Ventures, [n. d.]), and JP Morgan Chase has mentioned it in some capacity. There are also a handful of startups which claim to deal with fraud detection and identity. Of all the FinTech areas, this is the one with less apparent current interest.

There is a marked similarity between fraud detection and predictive policing. If the training set contains more examples of fraud committed by one user group, then the model will potentially act in a discriminatory manner. The disproportionate representation could come from historically greater scrutiny paid to customers from that user group, or from a higher rate of prosecution for example. This is exactly a pitfall of predictive policing - past bias is reflected in the model’s behavior. In predictive policing, if police patrol certain neighborhoods more often, then they will be more likely to make arrests in those areas. Because they are more likely to make arrests, a prediction system may think that factors associated with those areas are more indicative of criminal behavior. Analogously, if the data being fed to the fraud detection algorithm is based on looking at a particular set of characteristics (which could even include location), then it will be more likely to identify fraudulent transactions that look like those transactions. This aspect of the problem can very easily be posed as one of the relationship between classifications and membership within a protected group. Indeed, dealing with a

training set that is known or suspected to be biased is a well-known problem in fairness in machine learning.

The analogy of predictive policing is truly apt, in that bias in how certain crimes are treated may be reflected in the model. If the sorts of illegal activity the bank monitors for is biased, then the model itself won't be biased, but will have a biased outcome. As noted by Bolton and Hand, identifying money laundering requires a different set of approaches than identifying payment fraud. Since poor people will generally not have money to launder, let us assume that poorer people tend to commit more payment fraud. A system that checks for payment fraud but which does not check for money laundering will be biased because it will catch poor people but not wealthier people. The parallel to predictive policing here would be the existence of laws criminalizing public intoxication. A system that focuses only on those particular sorts of offenses is going to be biased because it ignores the broader factors that make particular people likely to commit those crimes as opposed to other kinds of crimes. A strict focus on the model's inputs and outputs may not reveal these kinds of bias.

3.3 Credit

Of all the possible applications of ML to FinTech, credit and risk assessment is probably the most well-studied from a theoretical perspective. In this setting, the bank or lender wants to decide whether or not to grant someone a loan or extend them credit, and if so on what terms. There are a number of startups which claim to address various aspects of this, along with some interest from Citigroup through investment in some other startups. The task is to determine (using some kind of data source) whether or not a particular applicant should receive credit and possibly, on what terms. Generally, the machine learning task will be to try and minimize the risk that the people who are granted loans will default.

There are a number of ways one can look at bias in credit decisions through the lens of machine learning. Joseph et al. (2016) looked at bias as being due to relative uncertainty about the credit-worthiness of applicants from minority groups. Either due to previous bias, or just due to random sampling, a bank may have fewer samples from members of a minority group, which could decrease their certainty when making predictions about members of that group. Or you could just have biased data. If financial institutions discriminated against a particular group in the past such that the group had worse outcomes, then a model that is trained on that data will be discriminatory. Regardless of the actual cause of the bias, the issue at root is the relationship between credit determinations and membership in a protected class.

However, despite the setting already being somewhat mathematized, there are ways in which the credit application scenario can evade the machine learning approach. Suppose that one way of increasing the banks' certainty that the decision to extend a loan is a good idea is by having someone who already owns a home co-sign your loan. If this is a true pattern, then a credit decision system should pick it up. However, this criteria will also produce bias because african-americans are less likely to own homes than caucasians (Chiteji and Stafford, 1999). Because the relationship in this scenario is "true", most fairness frameworks will see it as permissible. Strict group parity criteria would prevent this, but at the cost of less accurate prediction. While the machine learning approach can be used to model particular relationships, it can't provide an understanding of why a pattern might exist. That's not to say that a machine learning approach needs to be able to fully answer why patterns are the case. Rather, the governance process needs to be able to engage with and contextualize patterns that they see the model identifying. If you lack an understanding of why a particular relationship holds in the data, it will be very difficult to make decisions which are actually fair.

3.4 Customer Service

Chatbots are computer programs meant to interact in a conversational manner with customers. In FinTech, they are generally thought of in terms of providing customer service, either aiding or replacing human call center employees. A chatbot can range from using simple heuristics like keywords to more sophisticated methods like sentiment analysis or deep learning to determine how to respond to customer inquiries. Because customer service can be a large cost for a customer facing bank, there has been intense interest by large banking institutions. JP Morgan Chase recently hired a roboticist who specializes in human-computer interaction as their head of machine learning (Horowitz, 2018), and has also expressed interest in hiring people with experience in natural language processing

(NLP) at lower levels. Bank of America already has its own chatbot (Roberts, 2016), and Wells Fargo has expressed some interest as well (Wells Fargo, [n. d.]).

Because helpful behavior is such a contextually specific thing, this use case is especially illustrative of formalization bias. In fact, Friedman and Nissenbaum provide an example which very nearly fits this setting - saying a legal expert system advises defendants on whether or not to plea bargain by assuming that law can be spelled out in an unambiguous manner that is not subject to human and humane interpretations in context. A chatbot is in a similar position to their "legal expert system", in that its task is to help its user navigate a system in which it is not an expert. Similarly, if the chatbot is assisting users with functions involving the interpretation of bank policy, or other functions that are subject to human interpretation, then it will lead people astray. Users who are able to access and influence the human interpretation, possibly because they are able to spend the time and money it would take to reach a human representative, will not be subject to the strict interpretation of the chatbot.

If a chatbot doesn't understand vernacular or a dialect of a language, then the chatbot will be biased. The main benefit of a chatbot is that it can engage with customers in a conversational and hopefully easy to understand manner. If a group of users is forced to speak to the bot in an artificial or unnatural manner, then they are being denied those benefits. Understanding dialects and vernacular is very much a machine learning task as it has to do with how the model interprets and responds to speech. Indeed, prior work has looked at how a failure to accurately understand African-American Vernacular English (AAVE) could have serious implications for free speech (Duarte et al., 2017; Blodgett et al., 2016). Applied to this domain, similar issues arise.

3.5 Financial Advice

Predicting the behavior of certain financial quantities is an area of intense interest for many different purposes and parties (in some cases predating the term FinTech). The financial predictions can range from trying to predict particular stocks or markets, to predicting cash flow. In some cases the prediction task may be implicit in a maximization problem (that is, the algorithm tells the user what actions will maximize some goal function) or the raw prediction will be used in some manner. This task can be used for three different purposes: for market prediction, financial advising (either directly to a consumer, or as an aid to a professional advisor), and cash flow prediction. JP Morgan, Bank of America and Wells Fargo have all expressed some sorts of interest in it (though their interest in it does not appear to be consumer facing) (J.P. Morgan, [n. d.]; Verhage, 2017; Butcher, 2018). A few startups actually claim to use algorithmic methods to optimize investment portfolios in a way that does face the general public. This straddles the line between financial advising and market prediction.

Financial advice, though related to market or stock prediction is a distinct subcategory of ML based FinTech. Financial advising applications of ML may either be used to recommend investments to a user, or to try and get the user to change their behavior to meet particular financial goals. More rarely, the app will be a tool for a professional financial advisor. Generally, these apps will take information about the advisee's financial situation and goals, and produce some kind of recommendation. An even more specialized situation is cash-flow prediction. While more relevant to business applications, having an idea of expected account balance can be useful for freelancers and for payment processing institutions.

A key concern for people seeking to deploy machine learning systems in the real world is detecting population drift. That is when the assumptions that you made about the current conditions have shifted, causing the model to potentially perform much worse. A financial advising application which fails to assimilate new conditions and update the types of advice it provides potentially can lead to biased outcomes. If interest rates increase, and the bond market collapses, and the financial advising bot doesn't realize that the bond market is not a good place to invest, then when it recommends users buy bonds it will make a bad recommendation. If the recommendations are tailored to users, then it may provide bad advice to a particular group of users.

Another way in which a machine learning system can manifest bias in this system is if the recommendations the system makes are not actionable for a particular user group. If Iranian users are advised that the thing to do with their money is to invest in American securities, then the system is biased insofar as making those sorts of investments may be difficult. The actionability of particular

suggestions is outside the domain of a machine learning problem, or at least the financial advice machine learning problem.

4 Conclusion

Discrimination in machine learning systems is partially, but not totally, a machine learning problem. We have provided example tasks from the FinTech domain demonstrating how bias can be produced through a machine learning problem, and through broader factors of system design. Bias in a system that uses machine learning can arise from problems that can be understood in machine learning terms, and bias can also arise from other problems of value and design.

Not all problems of bias in machine learning are machine learning problems, but that doesn't mean that a model governance process should only look at only machine learning problems. Ultimately, the point of having a process to limit bias in a machine learning system is to prevent negative effects. Bias that does not arise directly from the machine learning problem still has undesirable effects in the world, and so should be considered within a process. This fact has implications for the sort of oversight a governance process needs to provide, and how a governance process should view fairness criteria and learning approaches. The governance process needs to be invoked throughout the system (not just model) design, implementation and use. The governance process should be seen as identifying and prioritizing areas of concern. The groups tasked with design and implementation then can treat the identified areas as design problems. In this view, the implementation team will pick and choose design elements as part of an attempt to respond to the areas that the governance process has identified. Some of these elements may involve understandings of bias as a machine learning problem, but others may not.

In order to develop these areas of concern, the governance process needs to consider the system in a holistic manner. We provide the following guiding questions for the governance process to ask when engaging in this.

- What should be the scope of governance process? As we have argued here, and others have argued elsewhere, computer systems don't end at the screen display. A governance process that is narrowly focused on model behavior will fail to fully account for potentially discriminatory behavior that the machine learning enables. Thus far we have emphasized taking a holistic view of the system, and indeed in determining the scope of the governance process the relevant organs should consider not just organizational actors (the model, the people developing the model, the people interpreting the model) but also how the goals of the project are determined and technical tradeoffs made. Obviously for some processes, there will be aspects that are out of the control of anyone within the organization (either because the goals are legally determined, or the entities cannot be influenced by the organization, or a governance process that would include that aspect of the system would be too unwieldy to function). Nevertheless, we suggest that the exclusion of a system aspect from scope should be a positive determination, i.e. that there are good reasons why the governance process should not cover that aspect.
- Having identified the relevant aspects, the determiners of the governance process should consider the sorts of interactions that the governance process should be concerned with. Interactions here referring both to the individual sorts of actions actors can take with respect to one another, as well as the broader patterns that could arise. This shouldn't be interpreted solely as interactions facilitated by the organization. Accessibility and actionability (exemplified by our analysis of customer service and of financial advice) are defined not by an absence of organizational actions, but by users not taking action. A myopic governance program may miss these phenomena if they are not specifically planned for.
- The two considerations considered above need to be guided by an understanding of the social and historical context in which the proposed system will operate. Without such an understanding, identifying actors and the types of problematic interactions will prove difficult. While this may sound burdensome, it need not be. A perfect understanding of any social phenomena is imperfect and bound up with the theory with which one approaches the problem. Consulting a multi-disciplinary array of domain experts is one way, another would be to consult with stake-holders. Indeed, this sort of research can be reused to help the designers of the system improve lower-level aspects of their system design.

The considerations contained here are high level and abstract. In that sense, to a technical audience, they are probably unsatisfying. It is our contention that defining these high level considerations will develop specific and concrete technical questions. As we have shown in the preceding sections, bias cannot be put in purely technical terms, so the strategy for addressing it cannot be purely technical. A governance process must first look across the system before looking into specific components.

References

Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (March 2018), 973–989. https://doi.org/10.1177/1461444816676645

angellist.com. [n. d.]. Startup Database. https://angel.co/companies

Tom Baker and Benedict Dellaert. 2017. Regulating Robo Advice across the Financial Services Industry Essay. *Iowa Law Review* 103 (2017), 713–750. https://heinonline.org/HOL/P?h=hein.journals/ilr103&i=729

Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California law review* 104 (2016), 671. https://doi.org/10.2139/ssrn.2477899

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. *arXiv:1608.08868 [cs]* (Aug. 2016). http://arxiv.org/abs/1608.08868 arXiv: 1608.08868.

Richard J. Bolton and David J. Hand. 2002. Statistical Fraud Detection: A Review. *Statist. Sci.* 17, 3 (2002), 235–249. https://www.jstor.org/stable/3182781

Andrew Burt, Brenda Leong, Stuart Shirrell, and Wang, George. 2018. *Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models*. Technical Report. https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf

Dan Butcher. 2018. BAML hires top machine-learning quant from J.P. Morgan. https://news.efinancialcareers.com/us-en/312526/baml-hires-top-machine-learning-quant-j-p-morgan

Ngina S Chiteji and Frank P Stafford. 1999. Portfolio Choices of Parents and Their Children as Young Adults: Asset Accumulation by African-American Families. *American Economic Review* 89, 2 (May 1999), 377–380. https://doi.org/10.1257/aer.89.2.377

Citi Ventures. [n. d.]. Portfolio - Citi Ventures. http://www.citi.com/ventures/portfolio.html

Natasha Duarte, (first), Llanso, Emma, and Loup, Anna. 2017. Mixed Messages? The Limits of Automated Social Media Content Analysis. https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, Cambridge, MA, 214–226. http://arxiv.org/abs/1104.3913 arXiv: 1104.3913.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (1996), 330–347.

S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer, and F. Giannotti. 2012. Injecting Discrimination and Privacy Awareness Into Pattern Discovery. In *2012 IEEE 12th International Conference on Data Mining Workshops*. 360–369. https://doi.org/10.1109/ICDMW.2012.51

Moritz Hardt, Eric Price, , and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

Julia Horowitz. 2018. JPMorgan's latest hire proves the bank is serious about artificial intelligence. https://money.cnn.com/2018/05/03/investing/jpmorgan-artificial-intelligence-chief/index.html

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc., Barcelona, Spain, 325–333. http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.pdf

J.P. Morgan. [n. d.]. Machine Learning | J.P. Morgan. https://www.jpmorgan.com/global/research/machine-learning

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 656–666. http://papers.nips.cc/paper/6668-avoiding-discrimination-through-causal-reasoning.pdf

Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (Jan. 2017), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Gustavo Ordóñez-Sanz. 2014. *Model Governance*. Technical Report. https://www.moodysanalytics.com/risk-perspectives-magazine/integrated-risk-management/approaches-to-implementation/model-governance

Deon Roberts. 2016. Bank of America unveils virtual assistant for mobile phones | Charlotte Observer. https://www.charlotteobserver.com/news/business/article110140572.html

Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. (2018), 15.

Julie Verhage. 2017. Wells Fargo Analysts Build the Robot That Could Take Their Jobs - Bloomberg. https://www.bloomberg.com/news/articles/2017-09-27/wells-fargo-research-analysts-invent-their-own-ai-replacement

Wells Fargo. [n. d.]. AI is making commerce smarter. http://welcome.wf.com/tech-banking/article/the-new-face-of-commerce-is-not-human.html